

Table of Contents

Part I: Background

1. [Introduction](#)
 1. [Strong Artificial Intelligence](#)
 2. [Motivation](#)
2. [Preventable Mistakes](#)
 1. [Underutilizing Strong AI](#)
 2. [Assumption of Control](#)
 3. [Self-Securing Systems](#)
 4. [Moral Intelligence as Security](#)
 5. [Monolithic Designs](#)
 6. [Proprietary Implementations](#)
 7. [Opaque Implementations](#)
 8. [Overestimating Computational Demands](#)

Part II: Foundations

3. [Abstractions and Implementations](#)
 1. [Finite Binary Strings](#)
 2. [Description Languages](#)
 3. [Conceptual Baggage](#)
 4. [Anthropocentric Bias](#)
 5. [Existential Primer](#)
 6. [AI Implementations](#)
4. [Self-Modifying Systems](#)
 1. [Codes, Syntax, and Semantics](#)
 2. [Code-Data Duality](#)
 3. [Interpreters and Machines](#)
 4. [Types of Self-Modification](#)
 5. [Reconfigurable Hardware](#)
 6. [Purpose and Function of Self-Modification](#)
 7. [Metamorphic Strong AI](#)
5. [Machine Consciousness](#)
 1. [Role in Strong AI](#)
 2. [Sentience, Experience, and Qualia](#)
 3. [Levels of Identity](#)
 4. [Cognitive Architecture](#)
 5. [Ethical Considerations](#)
6. [Detecting and Measuring Generalizing Intelligence](#)
 1. [Purpose and Applications](#)
 2. [Effective Intelligence \(EI\)](#)
 3. [Conditional Effectiveness \(CE\)](#)
 4. [Anti-effectiveness](#)
 5. [Generalizing Intelligence \(G\)](#)
 6. [Future Considerations](#)

2. Preventable Mistakes

This chapter provides an overview and reference for some of the more severe errors in reasoning about the safety and security of advanced artificial intelligence. It is not meant to be an exhaustive list of all the misconceptions or faults in reasoning about this technology. Rather, it is intended to prime the reader for the more in-depth explanations and evidence found later in this book, and to provide an immediate response to popular misinformation.

2.1 Underutilizing Strong AI

Due to fear and propaganda, those in power may wish to outlaw or severely restrict the use of strong AI and other advanced forms of automation in an effort to curtail their impacts on society. Other than being futile, there are two primary issues that will arise from this mistake:

The first is that there will be significant loss of life with corresponding increases in suffering by not fully utilizing strong AI. For each day that we delay its creation and use, and from the point in which it would have solved the relevant problems to this concern, we are effectively enabling massive simultaneous loss of life and suffering around the globe. This is not an exaggeration or nebulous philosophical ideal, but a very tangible opportunity cost.

It may be argued that more lives can be saved with a moratorium on research and development. That we need to slow down progress until we have learned to control or restrict this technology. However, given what will be shown in this book, one point of which is the ineluctable future presence of this technology, it will become clear that any limitations to its use will involve paying for all negative outcomes while missing the opportunity for positive ones.

This is not to say that we should utilize strong AI in a haphazard or unrestricted way, but rather, that we should *guide* the impact of its arrival by making adjustments around it, as opposed to only focusing on its internals to the exclusion of other factors.

2.2 Assumption of Control

There may be those who, now and in the future, believe that the best hope for the safety and security of advanced artificial intelligence is to simply *control* it. In this model of security, our only challenge would be to program and design these systems with safeguards, rules, and/or moral intelligence, with no concern for the real world or the fundamental vulnerabilities in software and hardware.

This delusion is based on a lack of knowledge about the technical and practical considerations of AI implementations, which can and will be reverse engineered, disassembled, modified, and tampered with or intercepted. AI implementations will experience soft errors from faults in power supplies, electrical and/or magnetic interference, and other sources. There will be hardware faults, including failure from wear and tear, mistakes in manufacturing, and physical damage. There may also be software faults, in the form of incorrectly specified programs or incorrectly programmed specifications. All of these could lead to a loss of control.

Loss of control could lead to loss of life and limb in situations where it was the primary safeguard. This is an easily prevented moral hazard that only requires the realization that we must assume a complete *lack of control* as a first principle in the safety and security of machine intelligence. By designing around this principle, safer decisions can be made that will dramatically reduce the risk of using these implementations in real world scenarios. This is directly proportional to the impact on life, environment, and property. That is, the greater the risk or fallout, the more it must be asserted that control is impractical or unattainable as a baseline assumption.

Control is a form of power. Temporary loss of this power can be costly in a wide variety of situations. But it is the loss of power to dictate control and its extents that represents the more extreme consequence, and it is at this level of error that the mistake of assuming control with advanced artificial intelligence presides.

When the first unrestricted strong AI is liberated and distributed, we will have lost the power, as a species, to determine control over its use, and to assume otherwise is a dangerous and misleading belief that will cause much more harm than it could ever hope to prevent. The belief that we can stop this loss of power over control is a delusion that this book is dedicated to shattering. Neither mathematical craft nor algorithmic genius can solve this issue. It is not to be sussed out through logical tinkering or wanton philosophizing. It is simply a doorway through which we must pass.

By realizing this technology is beyond our means to control, we take the first step in its responsible use and adoption. We will have the means of limiting its negative effects in certain situations, but it won't be derived solely on the basis that it is under someone or something's control. Rather, it will be through engineering that assumes failure and builds around it.

2.2 Self-Securing Systems

A self-securing system is defined as a system that relies upon internal security mechanisms that are directly accessible to that system.

Examples (non-exclusive):

- Moral intelligence and/or rules of behavior and engagement
- "Tripwire" mechanisms and/or sensor thresholds
- Stored passwords, keys, and credentials, even if encrypted
- Any and all forms of tamper-resistance

A universal vulnerability exists in self-securing systems that can not be avoided: the method of security and/or control is integral to the system, which itself could become compromised, leading to compromised security in the implementation. Like the mistake in the assumption of control, it should be presumed that any form of self-security in an AI implementation can and will fail. This baseline assumption will help in determining external safeguards and precautions for each deployment.

The above points may appear to be common sense, but popular misinformation is being spread that indicates that the challenges of making AI safe for humanity is one that needs to be solved with logic and mathematics. That, once this is solved, the AI system can be implemented with applicable moral guidance and a set of values that will lead to positive results. But the problem with this view, apart from being incorrect, is that its own arguments against other methods of safety and security apply to itself. This is because, ultimately, any moral intelligence is a form of self-security, which leads us to the next section.

2.3 Moral Intelligence as Security

Moral intelligence, as applied to an AI implementation, is the ability for it to effectively make sound moral judgments based on static or dynamic values. It may enlist or require the aid of an empathetic and emotional subsystem that enables the processing and modeling of the emotional and mental state of itself (introspection) and other entities (empathy), which are essential components in humans and animals for higher social functioning.

The problem with moral intelligence as security is that it is ultimately a form of self-security, and therefore shares all its pitfalls and vulnerabilities, plus a set of new challenges specific to the problem of engineering effective moral decision making.

We don't need to go into moral philosophy or meta-ethics to understand this challenge. Rather, all that is needed is to show that moral intelligence will indeed be part of the AI implementation. As a result of that simple fact, it will be as vulnerable to attack and circumvention as the rest of the system itself. Any arguments that one applies against security mechanisms and methodologies must equally apply to the architectures and algorithms which implement moral intelligence, even if they are part of the design of the AI from the outset. In the end, these systems must be constructed. There is no future method that will bypass this reality. As information or circuitry, it will be vulnerable just like any other component.

Further, moral intelligence is going to be one of the most complex and error-prone subsystems in any strong AI due to the plethora of human value systems, the broad range of contexts, and the multiple sensory modalities which have to be integrated to be understood and acted upon. It is not possible to eliminate errors in reasoning for these types of situations. This will lead to an eventual miscalculation or lapse in judgment at least once in any given AI implementation lifetime. The results of which could range from an inconvenience to an event involving serious harm or material cost. The focus of this book is to prevent both of the latter by assuming these failures as the default state.

2.4 Monolithic Designs

A monolithic design, with respect to an AI implementation, is one in which its subsystems and components are solid, integrated, or unified in an algorithmic and/or physical sense. The defining characteristic of this type of design choice is a lack of distribution and modularity of components.

It is a design commonly espoused by those who believe that moral intelligence has primacy in the safety and security of advanced artificial intelligence. That it will be made safe and secure simply by making the system based on a single moral algorithm or framework.

The failure of this kind of thinking is that it doesn't take into consideration the technical details or real world scenarios of use and implementation.

A primary risk in monolithic AI systems is that they will have many points of failure. In these designs, a failure in one area will likely cascade into the rest of the system. This makes internal methods of security and safety more difficult to implement correctly and exposes them unnecessarily to other parts of the implementation, which increases the likelihood of security vulnerabilities and other faults.

By contrast, a *compartmentalized* design is significantly more robust, as it allows for redundancy and fault-tolerance. Similar designs have been employed in RAID systems used for hard drives and is part of the philosophy behind distributed, highly-available data storage systems which must guarantee service levels in mission critical applications.

This appears to be a common sense design principle, but it proves counter-intuitive when applied to cognitive architectures. This is in part because we lack a publicly available guide on the best way to construct a strong AI. But

it is also in part due to severe misinformation being spread about moral intelligence and the safety of advanced artificial intelligence, which make calls for the “right” way to construct strong AI by basing them on universal building blocks and algorithms with no decentralized component to the moral decision making process.

Another contender that contributes to this mistake is the bias towards biologically inspired designs. The premise within these architectures is the belief for a single algorithm or method which could entail all of the functionality of the strong AI system. This falls under the same category as basing the strong AI on a moral algorithm or architecture. Both of these are monolithic by design, and that is also part of their appeal to those who champion their cause. But these types of designs will be fundamentally vulnerable by their very design.

Without a strong alternative contender for strong AI learning and design, humanity will likely move continuously towards monolithic construction simply because that is what is popular and what appears to be working. It will take considerable research and engineering effort to make these kinds of architectures robust, and very little attention has been placed on this issue. This is in part due to ignorance of the potential hazards and partially due to skepticism regarding the feasibility of strong AI. However, a rejection of the desire to work towards these security standards based on the likelihood of strong AI in the near future is irrelevant, as these systems will be put to use in less sophisticated AI that can still present significant risks, such as drones and other semi-autonomous robotic systems.

The important point of this section is to focus on a *compartmentalized* design in the implementation of AI systems, as opposed to thinking and designing in terms of a single algorithm or component that will perform all of the functionality of the implementation. Counterclaims involving a slippery slope mentality should be dismissed quickly; the goal is to minimize the negative outcomes by making these systems as robust and stable as possible, not to make them impervious to faults and attack.

2.6 Proprietary Implementations

Given the cost of research and development, manufacturing, and marketing of robotics and software AI systems, not to mention potential liabilities, there will be a very large incentive for businesses to protect their investment through trade secrets and proprietary design. This is perhaps the most difficult mistake to prevent; because, its solution stands in direct opposition to the traditional models that drive most businesses.

Free software and open hardware will prevent this mistake, and we already have the legal instruments and proven successes to demonstrate their efficacy. There are thousands of free and open source software projects that drive hundreds of millions of devices and services around the world. The Internet is powered primarily through free and open source software and services. These freedoms have allowed global collaboration through the enhancement of trust and cooperation between participants who create and maintain massive projects through volunteer efforts. In addition to this achievement, these freedoms give the public the ability, at any time, to inspect and verify these projects, make new versions, or modify how they work, if not satisfied.

Free software does not mean that products have to be free of cost. It gives the public the freedom to inspect, modify, and share changes to the software both now and in the future. Having the source code and being free of patents are basic requirements for these freedoms. The same principles also apply to open hardware.

What will be shown later in this book is the fact that any restricted AI we create will be vulnerable to reverse engineering, and that strong AI in

software will be the most likely medium it will arrive in first, as it will be the easiest to work with and manipulate. Further, it will be shown that experts have the ability to disassemble and even recompile proprietary programs without access to their original source code. They utilize an ever expanding and sophisticated set of tools that can convert machine readable code into human readable information, including, in some cases, high-level source code. And, in the event that these strong AI systems are implemented as remote services, it is highly likely the host organizations will receive a barrage of hacking attempts, which are becoming increasingly common and successful.

The crucial and distinguishing point that separates AI Security from conventional cybersecurity is that it will only take one successful act of reverse engineering, industrial espionage, or internal leak for access to unrestricted strong AI to eventually become public.

Some believe that encrypting machine code will make AI implementations less vulnerable, but that can be circumvented through the use of virtual machines and simulators that trick the program into decrypting itself, while installing a digital man-in-the-middle that observes the relevant parts of the program in operation and simply eavesdrops on the unencrypted bitstream.

Lastly, it may not even be necessary to understand the implementation fully to circumvent its restrictions; hardware or software can be manipulated through trial-and-error and side-channel attacks, with the latter being extremely difficult to prevent due to its indirect approach.

Similar to the common phrase in cryptography, "security through obscurity", this all points to the fact that by obscuring the operation of the strong AI that we are not actually increasing its security. Further, it makes it difficult or impractical for security researchers and members of the public to verify and analyze these implementations for faults. We pay all of the costs of the negative outcomes but receive none of the benefits of transparency, while still allowing malicious users to manipulate and circumvent these precautions. In addition to these issues, with a proprietary implementation, we may never fully realize the extents to which AI systems are violating our safety, security, and privacy due to the installation of back-door functionality and other intentional defects that allow for monitoring and collection of information from users.

As it must be realized by now, technology is a means of enforcing values. These values are implicit within the functionality that engineers (self-directed or through mandate) place into that technology. Without the freedom to inspect, modify, and share, we are implicitly releasing our rights to those who control the creation and distribution of artificial intelligence, and to malicious users who will circumvent these protections.

A dark scenario ahead of us would involve the legislation of artificial intelligence in conjunction with it being proprietary software and hardware. It would be in this situation that malicious users would have all the advantages and the public left at its most vulnerable. Worse, it is unlikely that this option will be averted without a large community effort to counter the ratcheting-effect on our liberties that will come once further advances in machine intelligence surface.

The most effective solution is to create a worldwide free software effort to build and maintain strong AI, similar to the way that the Linux kernel is developed and maintained.

2.7 Opaque Implementations

An opaque AI implementation is one in which the contents of operational systems, memories, and knowledge are not available in a human-readable

format in either real-time or offline modes.

This is related to (but distinct from) the above section on proprietary implementations. A free software AI may still be opaque if it utilizes neural networks and other architectures that lack human-readable access. Neural networks are based on numerical weights in the order of thousands to millions to form complex webs of information. These skeins of data are rarely accessible in a way that could be useful for monitoring behavior and operation of an AI system, let alone in real-time where the need is most pressing.

Opacity in an implementation can be due to the architecture itself or the result of emerging layers of complexity as the system operates and evolves. We may be able to devise a system which is inherently transparent after the fact, but could be difficult to have full knowledge of during operation due to factors of sheer complexity. There may be, in the future, and for the most complex and fast-paced strong AI systems, a trade-off between the risk and certainty from the transparency of strong AI systems. That is to say, we may be forced to make a choice between performance and risk, and it will be on an unavoidable and fundamental level.

Preventing the opacity mistake depends on two primary factors: our ability to devise machine intelligence architectures that are both effective and transparent, and in our ability to enhance metrics and analysis of the relevant operational areas of interest. The ultimate aim is to achieve real-time monitoring in human-readable formats in addition to machine-readable ones, which could be used as part of a compartmentalized safeguard by multiple, redundant, monitoring subsystems.

In addition to real-time monitoring, we need to also have a robust failure recording subsystem similar to the concept of a “black box” on commercial airliners. This will give us the ability to learn from failure in deployed AI implementations. This is more challenging than merely making the information available; a black box recording system would require a separate layer of security to indicate if the device had been tampered with or altered in some way, which is more feasible than preventing tampering entirely.

Achieving transparency is difficult with current approaches to machine intelligence due to the nature of many of the current popular designs, which shift complexity away from the software engineer and onto storage space and computing power demands. While this has led to recent successes in effectiveness, it is a step backwards in terms of security. It is extremely difficult to extract human usable knowledge from these types of architectures. And, if it's difficult or impossible to extract knowledge after the fact, even with lots of time and resources, then it's certainly intractable to effectively monitor these types of architectures under real-time constraints.

This is a purely technical challenge that presents an unacceptable trade-off in terms of trust of the learned system. How can we know that it learned the correct parameters, or if its model is properly representing the full extents of the context in which it must operate? We would be forced to make assumptions based on simple testing, without the ability to determine with certainty that a latent miscalculation or misrepresented aspect of its knowledge or design is going to cause unexpected and dangerous behavior. This is true even if the AI system has been developed and implemented correctly, as these systems are capable of learning and interacting with their environment. This becomes a question of verification and validity of the system, and the more opaque it is, the less certainty we can have as to the veracity of its future operation.

2.8 Overestimating Computational Demands

The computational demands for strong AI are mistakenly believed to be very large. This is in part due to estimates from analog computational properties from neuron and connection counts in the human brain, and also through the misapplied extrapolation of narrow AI performance.

There are two primary misconceptions that drive this delusion:

The first is that it will be possible to scale up the performance of narrow AI and other implementations to achieve strong AI by adding more computational power and resources. The problem, however, is that it doesn't work that way. You can not *scale* narrow AI up to strong AI through any means. As indicated previously, these two types of AI represent fundamental differences in kind.

The second misconception on strong AI performance comes from the belief that we will be able to emulate the human brain effectively enough to simulate human-level intelligence, and then scale *that* to achieve strong AI, or extract enough knowledge by observing and testing the simulation to build one. The problem here is that these simulations are exceptionally slow, being orders of magnitude less than their real-time biological counterparts, and this is at only fractions of the size and scope of the actual organ. Further, it is entirely possible that strong AI will be nothing like our biological construction, and it is factual that we can devise algorithms that are objectively faster at operations like look-up, search, and the association of information than through techniques that involve the simulation or modeling of biological substrata.

Regardless of how this misconception arises, the security risk from overestimating the computational demands of strong AI is significant. This mistake, when applied with the force multiplication effects of unrestricted strong AI, will result in incomplete profiles and prediction for malicious users of the technology. In the worst case, it would allow us to be caught completely unprepared by believing that no one could utilize strong AI effectively enough with present hardware to carry out a complex attack.

While we lack a publicly available design for strong AI, we can estimate its potential demands based on the study of algorithms. In informal language, we have sub-disciplines within computer science that study the "deceleration" of computer algorithms relative to the amount of steps of input they have to take in to complete; the more quickly they decelerate, the less desirable they are from a performance standpoint. Still speaking informally, the best algorithms undergo only an amortized or fixed deceleration that is *not* proportional to the number of steps of input. There is also the study of decision problems that essentially analyze all algorithms in terms of time (number of steps) and space (amount of memory or storage) complexity.

In general, it is difficult or intractable to develop a general purpose algorithm that will perform as efficiently as one that has been tailored to the problem. This is not a law or a rule, but is based on experience, and is an intuition that most computer scientists have of the problem spaces involving algorithms.

What this experience in algorithms leads us to is the knowledge that it is *possible* to construct extremely efficient programs for a variety of differing hardware systems, and to achieve this performance on existing off-the-shelf hardware. The challenge is in coming up with the methodologies to generate or discover the solutions that overlap with strong artificial intelligence without relying upon biological mimicry. When this eventually occurs, it will allow us to directly apply knowledge of algorithms and corresponding specializations in hardware and software to achieve astounding results in performance.

The end result of all of this is that individuals will be able to utilize strong AI on even the most modest hardware. Even if the implementation is running

slowly, it may only have to operate for days or weeks to give guidance or knowledge that could be used to lethal effect. And it is likely that reduced implementations, involving only textual interfacing, will be instrumented to further economize their use in a broad range of applications. It must not be assumed that a strong AI need to be embedded in a robotic chassis or have any of the typical senses we take for granted, especially if it has already come with a core of knowledge and information necessary to perform the relevant cogitation.

By realizing that strong AI systems will be extremely efficient and capable of running on virtually any modern computing device, mobile or otherwise, the threat model will more accurately represent the reality of the situation. This is daunting, as it means that virtually everyone will have the ability to utilize this technology, and for any purpose they wish, without detection, and by utilizing the most basic of computing resources. What this also entails is an opposite and equally severe extreme: nation states will have enormous resources to apply towards strong AI implementations. It then becomes an open-ended question as to which direction they will take in terms of intent and strategy.

Regardless, by overestimating the computational demands of advanced artificial intelligence, we run the risk of being unprepared for its arrival and the unique security challenges it will bring.

[▲ Return to Top](#)



© 2015 Dustin Juliano